Convolutive Demixing with Sparse Discrete Prior Models for Markov Sources

Radu Balan[1] and Justinian Rosca[1]

Siemens Corporate Research,
755 College Road East,
Princeton, NJ 08540
{radu.balan, justinian.rosca}@siemens.com

Abstract. In this paper we present a new source separation method based on dynamic sparse source signal models. Source signals are modeled in frequency domain as a product of a Bernoulli selection variable with a deterministic but unknown spectral amplitude. The Bernoulli variables are modeled in turn by first order Markov processes with transition probabilities learned from a training database. We consider a scenario where the mixing parameters are estimated by calibration. We obtain the MAP signal estimators and show they are implemented by a Viterbi decoding scheme. We validate this approach by simulations using TIMIT database, and compare the separation performance of this algorithm with our previous extended DUET method.

1 Introduction

Signal Separation is a well studied topic in signal processing. Many studies were published during the past 10 years, each of them considering the separation problem from different points of view. Once can use model complexity to classify these studies into four categories:

1. Simple models for both sources and mixing. Typical signals are modeled as independent random variables, in their original domain, or transformed domain (e.g. frequency domain). The mixing model is either instantaneous, or anechoic. The ICA problem [1], DUET algorithm ([2]), or [3] belong to this category;
2. Complex source models, but simple mixing models. An example of this type is separation of two speech signals from one recording using one microphone. In this case, source signals are modeled using complex stochastic models , e.g. AR processes in [4], HMMs in [5], or generalized exponentials in [6];
3. Complex mixing models, but simple source models. This is the case of standard convolutive ICA. For instance source signals are i.i.d. but the mixing operator is composed of unknown transfer functions. Thus the problem turns into a blind channel estimation as in e.g. [7-9];
4. Complex mixing and source models. For instance [10] uses AR to model source signals, and FIR transfer functions for mixing.

J. Rosca et al. (Eds.): ICA 2006, LNCS 3889, pp. 544–551, 2006.

We chose the complexity criterion in order to point out the basic trade-off of signal separation algorithms. A more complex mixing or source model may yield a better performance provided it fits well the data. However more complex models are less robust to mismatches than a simpler model, and may perform unexpectedly worse on real world data. In our prior experiments [11] we found that simple signal and mixing models yield surprisingly good results on real world data. Robustness to model uncertainties explains these good results. Indeed this is the case with DUET. The basic idea of the DUET approach is the assumption that for any time-frequency point, only one signal from the ensemble of source signals would use that time-frequency point. In [12] we extended this assumption in a system with D sensors to what we called *generalized W-disjoint orthogonality hypothesis* by allowing up to $D-1$ source signals to use simultaneously any time-frequency point. In both cases source signals were assumed mutually independent across both time and frequency. In other words, any two different time-frequency coefficients of the same source are assumed independent. However we would like to increase the power of source separation particularly when there exists prior knowledge about the sources (see also [5,6,13]). In this paper we propose an incremental increase in source model complexity combined with simple mixing model that conforms to our basic belief that models should not be more complicated than what is really needed in order to solve the problem. For this we allow for statistical dependencies of source signals across time. More precisely [14] postulates a signal model that states that the time-frequency coefficient $S(k,\omega)$ of a (speech) signal $s(t)$ factors as a product of a continuous random variable, say $G(k,\omega)$, and a 0/1 Bernoulli $b(k,\omega)$, $S(k,\omega)=b(k,\omega)G(k,\omega)$. This formula models sparse signals. See also [15] for a similar signal model. Denoting by q the probability of b to be 1, and by $p(\cdot)$ the p.d.f. of G, the p.d.f. of S turns into $p_S(S)=qp(S)+(1-q)\delta(S)$, with δ, the Dirac distribution. For L independent signals $S_1,\ldots,S_L$, the joint p.d.f. is obtained by conditioning with respect to the Bernoulli random variables. The rank k term, $0\leq k\leq N$, is associated to a case when exactly k sources are active, and the rest are zero. In [12] we showed that by truncating to the first N+1 terms the approximated joint p.d.f. corresponds to the case when *at most N sources are active simultaneously*, which constitutes the *generalized W-disjoint hypothesis*. This paper extends the signal model introduced before by assuming the Bernoulli variables are generated by a Markov process, while the complex amplitudes $G(k,\omega)$ are modeled as unknown deterministic variables. The application we target is a meeting transcription system (see Figure 1) where an array of microphones records the meeting, and the convolutive mixing parameters are learned during an initial calibration phase. Section 3 describes the statistical signal estimators. We show that signal estimation is similar to a Viterbi decoding scheme. Section 4 presents the methods for learning the transition probabilities of source models, and of the mixing parameters. Section 5 contains numerical results, and is followed by the conclusion section.

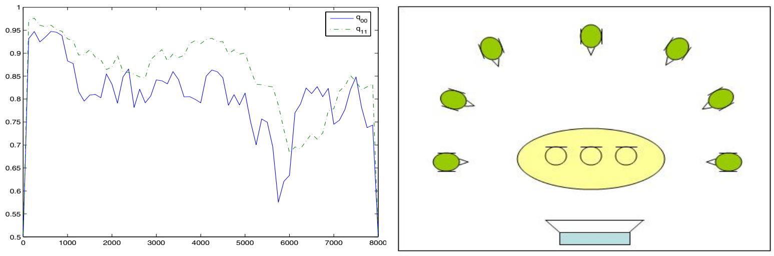

Fig. 1. Transition probabilitites of one signal for $\tau = 0.1$ (left plot), and the experimental setup (right plot)

2 Signal and Mixing Models

2.1 Convolutive Mixing Model

Consider the measurements of L source signals by an array of D sensors. In time domain the mixing model is $x_d(t) = \sum_{l=1}^{L} h_{d,l} \star s_l(t) + n_d(t)$, $1 \leq d \leq D$ where $n_1, \ldots, n_D$ are sensor noises, and $h_{d,l}$ are the impulse responses from source l to sensor d. We renormalize the sources by absorbing $h_{1,l}$ into the definition of source s_l.

We denote by $X_d(k,\omega)$, $S_l(k,\omega)$, $N_d(k,\omega)$ the short-time Fourier transform of signals $x_d(t), s_l(t)$, and $n_d(t)$, respectively, with respect to a window $W(t)$, where k is the frame index, and ω the frequency index. Then the convolutive mixing model turns into $X_d(k,\omega) = \sum_{l=1}^{L} A_{d,l}(\omega) S_l(k,\omega) + N_d(k,\omega)$. When no danger of confusion arises, we drop the arguments k, ω in X_d, S_l and N_d.

2.2 Signal Model

Consider a source signal $s(t)$, $1 \leq t \leq T$, and its associated short-time Fourier transform $S(k,\omega)$, $1 \leq k \leq K_{max}$, $0 \leq \omega \leq \Omega$. Each time-frequency coefficient $S(k,\omega)$ is modeled by the product $b(k,\omega)G(k,\omega)$ as before, where b is a Bernoulli (0/1) random variable, and G is an unknown deterministic complex amplitude. In previous work we assumed $\{b(k,\omega)\ ;\ k,\omega\}$ is a set of independent random variables. In this paper we preserve independence along the frequency index, but we introduce a Markov dependence along the time index. The independence in frequency is supported by the remark that local stationarity in time domain implies decorrelation of frequency components. Along the time index, our assumption amounts to $P(b(k,\omega)|b(k-1,\omega), b(k-2,\omega), \ldots, b(1,\omega)) = P(b(k,\omega)|b(k-1,\omega)) = \pi_\omega(b(k,\omega), b(k-1,\omega))$ where $\{\pi_\omega\}$ is the set of 2×2 matrices of probabilities of transition. By successive conditioning we obtain that: $P(\{b(k,\omega)\ ;\ k,\omega\}) = \prod_\omega P(b(1,\omega)) \prod_{k=2}^{K_{max}} \pi_\omega(b(k,\omega), b(k-1,\omega))$. For each source in the mixture we assume we have a database of training signals where we learn the matrices of transition probabilities and the set of initial probabilities (see Section 5).

For a collection of L source signals, we assume that only N Bernoulli variables are nonzero; the rest are zero. We denote by $\{(b_l(k,\omega))_{1\leq l\leq L}; k,\omega\}$ the collection of Bernoulli random variables, $\sigma(k,\omega)=\{l\ ;\ b_l(k,\omega)=1\}$ the N-set of nonzero components of $S(k,\omega)$, $(\pi^l_\omega)_{1\leq l\leq L,0\leq\omega\leq\Omega}$ the collection of transition probability matrices, $(P^l_\omega)_{1\leq l\leq L,0\leq\omega\leq\Omega}$ the collection of initial probabilities. Then the joint pdf becomes:

$$P(\{b_l(k,\omega)\ ;\ l,k,\omega\})=\prod_\omega Q^0_\omega(\sigma(1,\omega))\prod_{k\geq 2}Q_\omega(\sigma(k,\omega),\sigma(k-1,\omega))$$

where $Q_\omega(\sigma(k,\omega),\sigma(k-1,\omega))=\prod_{l=1}^L\pi^l_\omega(b_l(k,\omega),b_l(k-1,\omega))$, $Q^0_\omega(\sigma(1,\omega))=\prod_{l=1}^L P^l_\omega(b_l(1,\omega))$. The collection of all subsets $\sigma(k,\omega)$ defines a trajectory through the selection space S^N_L, the set of N-subsets of $\{1,2,\ldots,L\}$. Thus for each frequency ω we associate $\Sigma_\omega=\{\sigma(k,\omega)\ ;\ 1\leq k\leq K_{max}\}$ the selection space trajectory. Source estimation is then equivalent to estimating both the selection space trajectories $(\Sigma_\omega)_\omega$ and the complex amplitudes $\{G_l(k,\omega)\ ;\ l\in\sigma(k,\omega)\}$.

In this paper we assume that the mixing model is given by a convolutive mixture, signals $S_l(k,\omega)$ satisfy the signal model above, and noise components $N_d(k,\omega)$ are Gaussian i.i.d. with zero mean and spectral variance σ^2.

Our problem is: Estimate the source signals $(s_1(t),\ \ldots,\ s_L(t))_{1\leq t\leq T}$ given measurements $(x_1(t),\ \ldots,\ x_D(t))_{1\leq t\leq T}$ of the linear convolutive mixing model, and assuming the following:

1. The mixing matrix $A=(A_{d,l}(\omega))_{1\leq d\leq D,1\leq l\leq L}$ is known;
2. The noise $\{n(t)\}$ is i.i.d Gaussian with zero mean and known spectral power σ^2;
3. The components of signal S are independent and satisfy the stochastic model presented before, with known probabilities of transition $(\pi^l_\omega)_{l,\omega}$ and initial probabilities P^l_ω;
4. At every time-frequency point (k,ω) at most N components of $S(k,\omega)$ are non-zero, and N is known.

3 MAP Signal Estimation

In this paper we estimate the signals $(s_l(t))_{l,t}$ by maximizing the posterior distribution of the Bernoulli variables, and the likelihood of the complex amplitudes. Alternatively, using a uniform prior model on the amplitudes, our solution is a MAP estimator of both the selection variables and the complex amplitudes. The criterion to maximize is:

$$\begin{aligned}I=\prod_\omega &P(\{X(k,\omega);1\leq k\leq K_{max}\}|\{b_l(k,\omega),G(k,\omega);l,1\leq k\leq K_{max}\})\\ &\times P(\{b_l(k,\omega);l,1\leq k\leq K_{max}\})\end{aligned}\qquad(1)$$

We replace the Bernoulli variables by the set-valued variables $\Sigma_\omega=(\sigma(k,\omega))_{k,\omega}$, and we consider the reduced complex amplitude N-vector $\mathbf{G}_r(k,\omega)$ corresponding to nonzero components of S (in turn selected by $\sigma(k,\omega)$). We let $A_r(k,\omega)$

denote the $D \times N$ mixing matrix whose columns corresponds to the nonzero components of $S(k,\omega)$: $(A_r(k,\omega))_{d,m} = A_{d,l(m)}(\omega)$, where $l(m)$ is the m^{th} element of $\sigma(k,\omega)$. The first term decouples into a product of likelihoods at each time k; the second term is estimated before. Putting these two expressions together, the criterion to maximize becomes (up to a multiplicative constant term):

$$I(\Sigma, \mathbf{G}_r) = \prod_{\omega} \left[\prod_{k} exp\{-\frac{1}{\sigma^2} \|X - A_r \mathbf{G}_r\|^2\} \right] \times \left[\prod_{k \geq 2} Q_\omega(\sigma(k,\omega), \sigma(k-1,\omega)) \right] Q^0_\omega(\sigma(1,\omega))$$

Given $\sigma(k,\omega)$, at every (k,ω) we can solve for $\mathbf{G}_r(k,\omega)$ and obtain $\mathbf{G}_r(k,\omega) = (A_r^* A_r)^{-1} A_r^* X$. Taking the logarithm, flipping the sign, ignoring some constants, and replacing $\mathbf{G}_r$ by the above estimate, we obtain the following optimization problem

$$min_\Sigma \sum_\omega \sum_k [X^*(1 - A_r(A_r^* A_r)^{-1} A_r^*)X - \sigma^2 \log Q_\omega(\sigma(k,\omega), \sigma(k-1,\omega))] - \sigma^2 \log Q^0_\omega(\sigma(1,\omega))$$

Let us denote by

$$C(\sigma(k,\omega)) = X(k,\omega)^*(1 - A_r(k,\omega)(A_r^*(k,\omega)A_r(k,\omega))^{-1} A_r^*(k,\omega))X(k,\omega)$$

and

$$D(\sigma(k,\omega), \sigma(k-1,\omega)) = -\sigma^2 \log Q(\sigma(k,\omega), \sigma(k-1,\omega))$$

for $k \geq 2$. Then the optimization becomes

$$\min_{\Sigma_\omega} \sum_{k \geq 2} C(\sigma(k,\omega)) + D(\sigma(k,\omega), \sigma(k-1,\omega)) + C(\sigma(1,\omega) - Q^0_\omega(\sigma(1,\omega))$$

at every frequency ω. The solution represents a trajectory Σ_ω in the selection space $(S_L^N)^{K_{max}}$. The optimization can be efficiently implemented using a backward-forward best path propagation algorithm (Viterbi) widely used in channel decoding problems. The algorithm is as follows:

Algorithm

Step 1. (Initialization) Set $k = K_{max}$, and $J_k^*(s) = 0$ for all $s \in S_L^N$.

Step 2. (Backward propagation) For all $s \in S_L^N$ N-subsets of $\{1, 2, \ldots, L\}$ repeat

- For all $s' \in S_L^N$ compute $J(s,s') = J_k^*(s') + C(s') + D(s', s)$
- Find the minimum over s', and set $J_{k-1}^*(s) = min_{s'} J(s,s')$

Step 3. Decrement $k = k - 1$, and if $k > 1$ go Step 2.

Step 4. At $k = 1$, replace $C(s')$ by $C(s') - \sigma^2 \log Q^0_\omega(\sigma(1,\omega))$ and perform Step 2. Denote $\sigma^*(1,\omega) = argmin_s J_1^*(s)$.

Step 5. (Forward iteration) Set $k = 2$ and repeat until $k = K_{max}$:

- For all $s \in S_L^N$ compute $J(s) = C(s) + D(s, \sigma^*(k-1,\omega))$
- Find the minimum and set $\sigma^*(k,\omega) = argmin_s J(s)$
- Increment $k = k + 1$.

4 Model Training

4.1 Transition and Intial Probabities Estimation

For training we used a fixed sentence uttered by the corresponding speaker. We assumed the recorded voice is made of two components: one part which is critical to understanding, and a second component which can be removed losslessly from an information point of view. Thus $s = s_{critic} + s_{extra}$. Assuming the first component has a Laplace (or even peackier) distribution in frequency domain whereas the second component is Gaussian, the estimation of s_{critic} is done by (soft, or hard) thresholding of the measured signal. We chose a threshold proportional to square root of signal spectral power. Thus, in case of hard thresholding $S_{critical}(k,\omega) = S(k,\omega)$ if $|S(k,\omega)| \geq \tau\sqrt{R_s(\omega)}$, and is zero otherwise. The factor τ is chosen so that the thresholded signal sounds almost identical to the original signal s. Subjective experimentation showed that a factor $\tau = 0.1$ satisfies this requirement. Once $\{S_{critical}(k,\omega); k,\omega\}$ has been obtained, we estimate the binary sequence $\{b(k,\omega); k,\omega\}$ simply by setting $b(k,\omega) = 1$ for $S_{critical}(k,\omega) \neq 0$, and 0 otherwise. From the binary sequence $\{b(k,\omega); k,\omega\}$ we estimate the transition probability matrices π_ω and initial probabilitites P_ω by maximum likelihood estimators: $\pi_\omega(0,0) = \frac{N_{00}}{N_{00}+N_{01}}$, $\pi_\omega(1,0) = 1 - \pi_\omega(0,0)$, $\pi_\omega(1,1) = \frac{N_{11}}{N_{10}+N_{11}}$, $\pi_\omega(0,1) = 1 - \pi_\omega(1,1)$, $P_\omega(1) = \frac{N_1}{N_0+N_1}$, $P_\omega(0) = 1 - P_\omega(1)$, where N_0, N_1, N_{00}, N_{01}, N_{10}, N_{11} are, respectively, the number of 0's, 1's, 00's, 01's, 10's, 11's in the binary training sequence $(b(k,\omega))_k$. Figure 1 plots an example of the distributions $\pi_\omega(0,0)$ and $\pi_\omega(1,1)$.

4.2 Mixing Parameters Estimation

Consider the case one source only is active. Then the frequency representation of the recorded signal turns into $X(k,\omega) = a(\omega)S(k,\omega) + N(k,\omega)$, where $a(\omega)$ is the "steering vector" associated to source S. We use the maximum likelihood estimation to estimate a. Assuming Gaussian i.i.d. noise, the resulting maximum likelihood estimator yields $\hat{a}(\omega)$ the eigenvector corresponding to the largest eigenvalue of the sampled covariance matrix, normalized so that $\hat{a}_1 = 1$, $R\hat{a} = \lambda\hat{a}$, $R(\omega) = \sum_k X(k,\omega)X^*(k,\omega)$.

5 Experimental Evaluation

Consider the setup of a meeting recording system as depicted in Figure 1: $L = 7$ speakers placed around a conference table are recorded by a video camera (for eventual postprocessing) and an array of microphones. During the calibration phase both the source model parameters and the mixing parameters were learned. In our simulations we used a linear array with inter-microphone distance $d_a = 5$ cm and sampling frequency $f_s = 16$ KHz. The simulated mixing environment was weakly echoic with a reverberation time below 10ms. We used 4 female and 3 male speakers from the TIMIT database at positions located at multiple of 30 degrees. Testing was done on wavefiles of around 10 seconds of normal speech. We added Gaussian noise with $\sigma = 0.1$ (note σ is an absolute value rather than relative to signals). We tested for $N = 1$ and $N = 2$ (the number of simultaneous speakers), even though all $L = 7$ speakers were active most of the time. We estimated each source using the MAP-based Estimation Algorithm presented in Section 4 for four choices of priors: 1) use the initial distribution and transition probabilities learned from the training database as presented before; 2) use uniform initial distribution probabilities but the transition probabilities learned from the training database; 3) use uniform transition probabilities, but initial probabilities learned from the training database; 4) use uniform distributions for both the initial distribution and for the transition probabilities. This last combination of priors turns our MAP algorithm into the extended DUET presented in [12].

We compared these algorithms with respect to the Signal To Interference Plus Noise Ratio (SINR) Gain. The SINR gain for component l is defined by:

$$SINRg_l = oSINR - iSINR = 10\log_{10}\frac{E(x_1 - s_l)}{E(\hat{s}_l - s_l)}$$

where $E(z)$ is the energy of signal z, and $x_1, s_l, \hat{s}_l$ are respectively, the microphone 1 measured signal, input signal l at microphone 1, and the l^{th} estimated signal. The larger the $SINRg$ the better. We experimentally verified that the choice for initial distribution probabilities does not have almost any effect on the outputs. In Figure 2 we plot the SINR gain as function of number of microphones D, for our setup with $L = 7$ sources, and a variable number of microphones ranging from 2 to 6, for two hypotheses: $N = 1$ and $N = 2$, respectively. We notice the gain is an increasing function of number of microphones, and our MAP algorithm (called Markov, in Figure) outperforms DUET by about 1 dB in average.

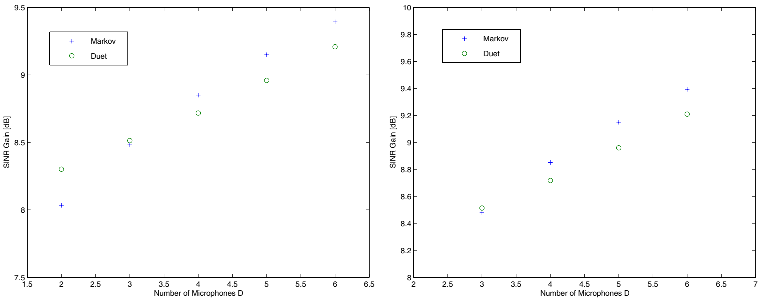

Fig. 2. SINR Gain for $N = 1$ (left plot) and $N = 2$ (right plot), for $L = 7$ sources and a variable array ranging from 2 to 6 microphones

6 Conclusions

In this paper we presented a novel signal separation algorithm that extends our past DUET algorithm. The algorithms works for underdetermined cases, when there are fewer sensors than sources, and in the presence of noise. The main assumptions are: (i) source signals have sparse time-frequency representations (although another representation, such as time-scale, would work as well); (ii) each frequency is independent from one another; (iii) the binary selection variables obey a homogeneous Markov process model, with transition and initial probabilities learned from a training database. We derived the MAP estimator of binary selection variables and ML of the complex signal TF coefficients, and show it can be efficiently implemented using a Viterbi decoding scheme. Next we validated our solution in a 7-voice, and 2 to 6 calibrated microphone array setup. We obtained an improvement of about 1 dB compared with the previous DUET algorithm, and no noticeable distortions.

References

1. Pierre Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.
2. A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," in *Proc. ICASSP*, 2000.
3. M. Aoki, M. Okamoto, S. Aoki, and H. Matsui, "Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones," *Acoust. Sci. & Tech.*, vol. 22, no. 2, pp. 149–157, 2001.
4. R. Balan, A. Jourjine, and J Rosca, "Ar processes and sources can be reconstructed from degenerate mixtures," in *Proc. ICA*, 1999, pp. 467–472.
5. S. T. Roweis, "One microphone source separation," in *Neural Information Processing Systems 13 (NIPS)*, 2000, pp. 793–799.
6. G.J. Jang and T-W Lee, "A probabilistic approach to single channel blind signal separation," in *Proc. of NIPS*, 2002.
7. H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. SAP*, vol. 12, no. 5, pp. 530–538, 2004.
8. A. Hyvarinen, J. Karhunen, and E. Oja, *Independent component analysis*, John Wiley and Sons, 2001.
9. J.F. Cardoso, "Infomax and maximum likelihood for blind source separation," *IEEE Signal Processing Letters*, vol. 4, no. 4, pp. 112–114, April 1997.
10. E. Weinstein, A.V. Oppenheim, M. Feder, and J.R. Buck, "Iterative and sequential algorithms for multisensor signal enhancement," *IEEE Trans. on SP*, vol. 42, no. 4, pp. 846–859, 1994.
11. R. Balan, J. Rosca, and S. Rickard, "Robustness of parametric source demixing in echoic environments," in *Proc. ICA*, December 2001.
12. J. Rosca, C. Borss, and R. Balan, "Generalized sparse signal mixing model and application to noisy blind source separation," in *Proc. ICASSP*, 2004.
13. S. Hosseini, C. Jutten, and D. Pham, "Markovian source separation," *IEEE Trans. on Sig. Proc.*, vol. 51, pp. 3009–3019, 2003.
14. R. Balan and J. Rosca, "Statistical properties of STFT ratios for two channel systems and applications to blind source separation," in *Proc. ICA-BSS*, 2000.
15. P.J Wolfe, S.J. Godsill, and W.J. Ng, "Bayesian variable selection and regularization for time-frequency surface estimation," *J.R.Statist.Soc.B*, vol. 66, no. Part 3, pp. 575–589, 2004.